IncluSet: A Data Surfacing Repository for Accessibility Datasets

Hernisa Kacorri
University of Maryland, College Park
hernisa@umd.edu

Utkarsh Dwivedi
University of Maryland, College Park
udwivedi@terpmail.umd.edu

Sravya Amancherla
University of Maryland, College Park
sravyaa@umd.edu

Mayanka K. Jha
University of Maryland, College Park
mjha@umd.edu

Riya Chanduka
University of Maryland, College Park
chanduka@terpmail.umd.edu

ABSTRACT

Datasets and data sharing play an important role for innovation, benchmarking, mitigating bias, and understanding the complexity of real world AI-infused applications. However, there is a scarcity of available data generated by people with disabilities with the potential for training or evaluating machine learning models. This is partially due to smaller populations, disparate characteristics, lack of expertise for data annotation, as well as privacy concerns. Even when data are collected and are publicly available, it is often difficult to locate them. We present a novel data surfacing repository, called IncluSet, that allows researchers and the disability community to discover and link accessibility datasets. The repository is pre-populated with information about 139 existing datasets: 65 made publicly available, 25 available upon request, and 49 not shared by the authors but described in their manuscripts. More importantly, IncluSet is designed to expose existing and new dataset contributions so they may be discoverable through Google Dataset Search.

KEYWORDS

disability, dataset, repository, bias, artificial intelligence

ACM Reference Format:
Hernisa Kacorri, Utkarsh Dwivedi, Sravya Amancherla, Mayanka K. Jha, and Riya Chanduka. 2020. IncluSet: A Data Surfacing Repository for Accessibility Datasets. In *ASSETS '20: The 22nd International ACM SIGACCESS Conference on Computers and Accessibility, October 26–28, 2020, Athens, Greece.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3373625.3418026

1 INTRODUCTION

Data lie at the heart of innovative solutions leveraging advances in machine learning and artificial intelligence. They are used to train models, benchmark their performance, understand the complexity of real-world phenomena, and exclusion [18]– issues of fairness pertaining to underrepresented populations. Knowing the importance and cost of data collection and annotation as well as their potential benefit to the public, researchers, companies and governments often release data publicly or upon request [9]. However, sharing practices vary considerably among research communities and, contexts [21] with the focus in our work being accessibility.

Sharing has often served to attract, nurture, and challenge data scientists and technologists to work on specific problems. Many fields, including the health community [6, 33], have seized this opportunity to promote data science in their area. We observe the start of a similar trend in accessibility, *e.g.* the VizWiz data challenge [11] calling for the computer vision community to work on visual question answering problems that can serve people with visual impairments. However, scarcity of large datasets generated from people with disabilities that can be used in AI-infused technologies remain one of the field's biggest challenges [2]. While this is partly due to a smaller population [26], there are other factors specific to these user groups. People vary in their individual preferences and environments, but people with disabilities lend further dimensions with disparate characteristics, even within a given disability. Moreover, data annotation requires domain knowledge that few possess making it difficult to crowdsource annotations. For instance, creating datasets for emotion recognition for people who are Deaf[1] requires sign language fluency since, in addition to emotion, facial expressions are an inherent part of the language often used to convey syntax when signing [22]. And more importantly, there are privacy and ethical concerns for creating and sharing accessibility datasets[2] as people who have distinct data patterns may be more susceptible to data abuse and misuse [10, 13, 30].

In this paper, we present our efforts in creating Incluset, a publicly available repository, that can help the community discover and link to accessibility datasets that include data generated by people with disabilities and older adults that relate to technology; our focus is on datasets that have the potential for training machine learning models. We use the term `surfacing repository` to indicate that *none of the datasets are stored in our servers*, but metadata about where they can be found, the population represented, type of data and annotations, and technology used are. Currently, the repository is pre-populated with a total of 139 existing accessibility datasets that were manually located over 2018-2020 with few examples shown in Figure 1. A key challenge we identified in collecting and analyzing these accessibility datasets is that many are difficult to locate. They are spread across many venues and do not surface through a simple search. Moreover, they lack consistent descriptions and require manual screening. We believe that this repository does not only contribute to transparency but also moves us a step

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
ASSETS '20, October 26–28, 2020, Athens, Greece
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7103-2/20/10.
https://doi.org/10.1145/3373625.3418026

[1] Here we adopt the following convention: Deaf (capitalized) describing members of the linguistic community of sign language users.
[2] Accessibility datasets here refer to data that can be used to train machine learning models and that are generated from people with disabilities and older adults.

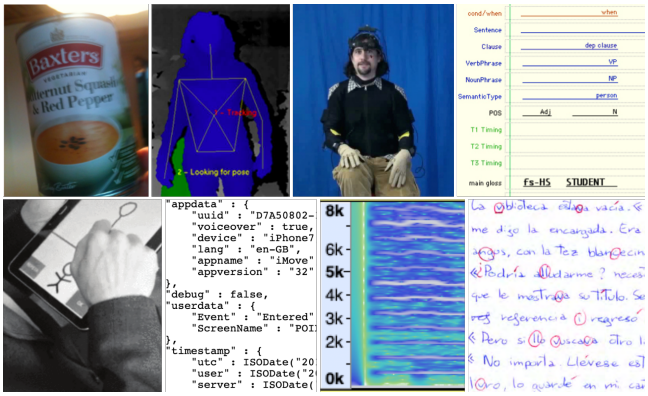

Figure 1: Examples of datasets generated by people with disability including from left to right photos taken by blind people [11] sign language videos and annotations [4, 19] stroke gestures by people with low vision [31], mobility app logs from people with visual impairments [17], audio recording from people with dysphonia [3], and text written by people with dyslexia [25].

towards better understanding local sharing standards in our community and potential concerns that can feed into the conversations to follow; we are currently completing our coding and analysis and hope to share the results in a follow-up publication. In this paper, we introduce IncluSet and its functionalities.

2 DISABILITY DATA

There have been many prior efforts on collecting and analyzing disability data. To our knowledge, they have mainly focused across dimensions such as longitudinal data on the demographics, diagnoses, causes of injury, interventions, outcomes, and costs (e.g., BMS National Database [5]). Previous efforts similar to this work have focused on collecting and analyzing survey data from people with disabilities [20], accessible websites [29], and geographical data on the accessibility of physical environments [7]. The proposed work is complementary to but qualitatively different from these prior efforts because it focuses on annotated data resources that include raw data generated by people with disabilities (*e.g.* extrasensory data such as accelerometer data and images taken by blind users) with some of them visualized in Figure 1. What's important (essential) about these data is that they can be used to drive innovative assistive technology or to benchmark models trained on data that do not include people with disabilities. For example, video recordings of Deaf signers with annotated facial expressions timestamps can be used to either train sign language avatars to be more understandable [16] or explore the performance of facial expression recognition technologies that might misread linguistically meaningful facial expressions during signing as emotions [27].

3 INCLUSET

IncluSet is a web application accessible at https://incluset.com/, launched in July 2020. ReactJS and NodeJS are used to develop the application. As shown in Figure 2, users can search for datasets that are already linked in the repository or sign in to link new datasets. Each dataset has an image[3], a title, dataset creators, a short description of the dataset, the year it was created/released, the number of people it was collected from, one ore more disability categories[4] describing the population of interest, and a list of data types. Users can reach the dataset through the direct link when available, read the paper where the data is described, or contact the dataset creators if contact information is available. As shown in Figure 2, out of the 139 datasets only 65 can be downloaded directly (*e.g.*, through a webpage available by the dataset creators), 25 are available upon request (*e.g.*, an email indicated by the creators), and 49 don't include any sharing information (we link to the papers). Additional tags are used to highlight terms originally used by the dataset creators such as description of the people who contributed to the data collection, technology used, fine-grained description of the data types, and purpose of the data collection. Last, users can upvote a dataset or see the upvote count[5].

3.1 Searching for Datasets

IncluSet enables different searching approaches. Users may explore the repository by prioritizing the dataset sharing strategies across the following tabs: `Download` for publicly available datasets that can be downloaded by anyone; `Request` for datasets that are available upon request from the dataset creators; and `Contact` for those that don't include sharing information. Users can further narrow they search by focusing on specific type of data such as audio, video, text, motion, image, logs, and sensing or by focusing on specific disability categories such as autism, cognitive, developmental, health, hearing, learning, mobility, speech, and vision. A search bar is being implemented to allow for direct navigation.

3.2 Adding New Datasets

Data sharing platforms vary across different venues. For example, the Language Resources and Evaluation Conference, where many sign language datasets can be found, has made the contribution to the LRE Map [1] mandatory since 2018. The map monitors the use and creation of language resources such as datasets and tools. Other venues use platforms such as Kaggle [15]. We observed that more often, links or request information for accessibility datasets were buried in some footnote or a specific section on the manuscript, making it challenging to discover. IncluSet provides an opportunity to researchers to increase the visibility of their datasets. In contrast to the UCI Machine Learning Repository [8], where dataset creators are called to donate their datasets, IncluSet only requires metadata about the datasets not the dataset itself; it merely points to any information about the dataset. Moreover, to increase its sustainability, IncluSet enables anyone that stumbles upon a related accessibility dataset to submit information a dataset. Thus, one can help surface datasets from others and remove the additional burden from the creators to report their dataset to one more repository. All "submitted" datasets are reviewed by moderators for completeness.

[3]When an image is not provided the website uses the first disability category to generate a default icon.
[4]The Individuals with Disabilities Education Act (IDEA) categorization is currently used though we are still iterating with the disability community.
[5]The upvote counts in Figure 2 were randomly generated to demonstrate functionality since IncluSet was recently launched.

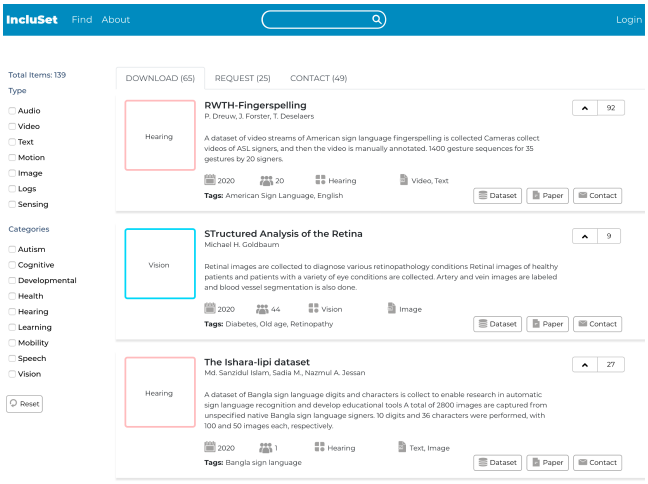

Figure 2: IncluSet: search for accessibility datasets directly or by filtering through data types, disability categories, and sharing strategies.

Users need to log in to add a dataset, so they can be contacted by the moderators and access the status of their dataset entries (e.g., whether it is published or additional information is needed). We believe that IncluSet can surface datasets and speed up their discovery as making datasets public through institutes [24] and funding agencies [28] tends to take longer [32].

3.3 Surfacing Datasets

Half-way through our dataset collection, we were excited to see Google deploy the new Dataset Search engine [23]. Unfortunately, at that time only 1 dataset related to accessibility, VizWiz [12] surfaced. To enable broader dataset discovery for accessibility, we implemented the Google Schema for all our datasets in the repository; through the react-schemaorg [14] we enable Google Search engine to crawl IncluSet and surface the datasets. We believe that IncluSet can contribute to innovation, benchmarking, mitigating bias, and understanding the complexity of real world AI-infused applications by promoting inclusion and accessibility while adding transparency to our data sharing practices.

ACKNOWLEDGMENTS

This work is supported by the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR), ACL, HHS (#90REGE0008). The opinions herein are those of the authors.

REFERENCES

[1] European Language Resources Association. 2017. LRE Map: Finding new ways through language resources. http://lremap.elra.info/

[2] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudrealt, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. *arXiv preprint arXiv:1908.08597* (2019).

[3] Ugo Cesari, Giuseppe De Pietro, Elio Marciano, Ciro Niri, Giovanna Sannino, and Laura Verde. 2018. A new database of healthy and pathological voices. *Computers & Electrical Engineering* 68 (May 2018), 310–321. https://doi.org/10.1016/j.compeleceng.2018.04.008

[4] Helen Cooper, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden. 2017. Sign Language Recognition Using Sub-units. In *Gesture Recognition*, Sergio Escalera, Isabelle Guyon, and Vassilis Athitsos (Eds.). Springer International Publishing, Cham, 89–118. https://doi.org/10.1007/978-3-319-57021-1_3

[5] Burn Model System National Data and Statistical Center. 1994. Burn Model System: Advancing recovery through knowledge. http://burndata.washington.edu/

[6] Carol C. Diamond, Farzad Mostashari, and Clay Shirky. 2009. Collecting And Sharing Data For Population Health: A New Paradigm. *Health Affairs* 28, 2 (2009), 454–466. https://doi.org/10.1377/hlthaff.28.2.454 arXiv:https://doi.org/10.1377/hlthaff.28.2.454

[7] Chaohai Ding, Mike Wald, and Gary Wills. 2014. A Survey of Open Accessibility Data. In *Proceedings of the 11th Web for All Conference* (Seoul, Korea) *(W4A '14)*. ACM, New York, NY, USA, Article 37, 4 pages. https://doi.org/10.1145/2596695.2596708

[8] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[9] Benedikt Fecher, Sascha Friesike, and Marcel Hebing. 2015. What Drives Academic Data Sharing? *PLOS ONE* 10, 2 (02 2015), 1–25. https://doi.org/10.1371/journal.pone.0118053

[10] Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, and Meredith Ringel Morris. 2019. Toward Fairness in AI for People with Disabilities: A Research Roadmap. *arXiv preprint arXiv:1907.02227* (2019).

[11] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Jun 2018). https://doi.org/10.1109/cvpr.2018.00380

[12] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3608–3617.

[13] Foad Hamidi, Kellie Poneres, Aaron Massey, and Amy Hurst. 2018. Who Should Have Access to My Pointing Data?: Privacy Tradeoffs of Adaptive Assistive Technologies. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (Galway, Ireland) *(ASSETS '18)*. ACM, New York, NY, USA, 203–216. https://doi.org/10.1145/3234695.3239331

[14] Google Inc. 2020. React-schemaorg. https://github.com/google/react-schemaorg

[15] Kaggle Inc. 2017. Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/

[16] Hernisa Kacorri and Matt Huenerfauth. 2016. Continuous Profile Models in ASL Syntactic Facial Expression Synthesis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 2084–2093. https://doi.org/10.18653/v1/P16-1196

[17] Hernisa Kacorri, Sergio Mascetti, Andrea Gerino, Dragan Ahmetovic, Hironobu Takagi, and Chieko Asakawa. 2016. Supporting Orientation of People with Visual Impairment: Analysis of Large Scale Usage Data. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility* (Reno, Nevada, USA) *(ASSETS '16)*. ACM, New York, NY, USA, 151–159. https://doi.org/10.1145/2982142.2982178

[18] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. 2017. Identifying Unknown Unknowns in the Open World: Representations and Policies for Guided Exploration. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (San Francisco, California, USA) *(AAAI'17)*. AAAI Press, 2124–2132. http://dl.acm.org/citation.cfm?id=3298483.3298546

[19] Pengfei Lu and Matt Huenerfauth. 2012. Cuny american sign language motion-capture corpus: first release. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, The 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey*.

[20] Jason Markesich. 2008. Surveying Persons with Disabilities: A Source Guide, Version II. (10 2008).

[21] Ingeborg Meijer, Stephane Berghmans, Helena Cousijn, Clifford Tatum, Gemma Deakin, Andrew Plume, Alex Rushforth, Adrian Mulligan, Sarah de Rijcke, Stacey Tobin, Thed Van Leeuwen, and Ludo Waltman. 2017. Open Data: the researcher perspective. (04 2017). https://doi.org/10.17632/bwrnfb4bvh.1

[22] Carol Jan Neidle, Judy Kegl, Benjamin Bahan, Dawn MacLaughlin, and Robert G Lee. 2000. *The syntax of American Sign Language: Functional categories and hierarchical structure*. MIT press.

[23] Natasha Noy, Matthew Burgess, and Dan Brickley. 2019. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *28th Web Conference (WebConf 2019)*.

[24] National Institute of Health (NIH). 2015. NIH Institutional certifications. https://osp.od.nih.gov/scientific-sharing/institutional-certifications/

[25] Luz Rello, Ricardo Baeza-Yates, and Joaquim Llisterri. 2014. DysList: An Annotated Resource of Dyslexic Errors. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Languages Resources Association (ELRA), Reykjavik, Iceland, 1289–1296. http://www.lrec-conf.org/proceedings/lrec2014/pdf/612_Paper.pdf

[26] Andrew Sears and Vicki L. Hanson. 2012. Representing Users in Accessibility Research. *ACM Trans. Access. Comput.* 4, 2, Article 7 (March 2012), 6 pages. https://doi.org/10.1145/2141943.2141945

[27] Irene Rogan Shaffer. 2018. Exploring the Performance of Facial Expression Recognition Technologies on Deaf Adults and Their Children. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (Galway, Ireland) *(ASSETS '18)*. ACM, New York, NY, USA, 474–476. https://doi.org/10.1145/3234695.3240986

[28] RF Terry, K Littler, and PL Olliaro. 2018. Sharing health research data ? the role of funders in improving the impact [version 1; peer review: 3 approved with reservations]. *F1000Research* 7, 1641 (2018). https://doi.org/10.12688/f1000research.16523.1

[29] Christian Thomsen and Torben Bach Pedersen. 2006. Building a Web Warehouse for Accessibility Data. In *Proceedings of the 9th ACM International Workshop on Data Warehousing and OLAP* (Arlington, Virginia, USA) *(DOLAP '06)*. ACM, New York, NY, USA, 43–50. https://doi.org/10.1145/1183512.1183522

[30] Jutta Treviranus. 2019. The Value of Being Different. In *Proceedings of the 16th Web For All 2019 Personalization - Personalizing the Web* (San Francisco, CA, USA) *(W4A '19)*. ACM, New York, NY, USA, Article 1, 7 pages. https://doi.org/10.1145/3315002.3332429

[31] R. Vatavu, B. Gheran, and M. D. Schipor. 2018. The Impact of Low Vision on Touch-Gesture Articulation on Mobile Devices. *IEEE Pervasive Computing* 17, 1 (Jan. 2018), 27–37. https://doi.org/10.1109/MPRV.2018.011591059

[32] Naomi Waithira, Brian Mutinda, and Phaik Yeong Cheah. 2019. Data management and sharing policy: the first step towards promoting data sharing. *BMC Medicine* 17, 1 (April 2019), 80. https://doi.org/10.1186/s12916-019-1315-8

[33] Mark Walport and Paul Brest. 2011. Sharing research data to improve public health. *The Lancet* 377, 9765 (2019/09/15 2011), 537–539. https://doi.org/10.1016/S0140-6736(10)62234-9